Identificação de padrões alimentares por regressão por redução de posto usando o programa SAS

Identification of dietary patterns with reduced rank regression using SAS software

Vanessa Schierholt da Silva[1], Vanessa Leotti Torman[1,2], Bianca Del Ponte da Silva[2], Marilda Neutzling[2], Maria Teresa Anselmo Olinto[3,4], Suzi Alves Camey[1,2]

Revista HCPA;31(4):497-511

[1]Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul, UFRGS.

[2]Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, UFRGS.

[3]Programa de Pós-Graduação em Saúde Coletiva, Universidade do Vale do Rio dos Sinos, UNISINOS.

[4]Departmento de Nutrição, Universidade Federal de Ciências da Saúde de Porto Alegre.

Contato
Suzi Camey.
camey@mat.ufrgs.br
Porto Alegre, RS, Brasil

Resumo

Introdução: a regressão por redução de posto (RRR) é uma técnica que vem sendo empregada na epidemiologia nutricional desde 2004. O objetivo dela é encontrar padrões alimentares associados a algum desfecho. Assim, ela é considerada uma técnica que combina informações a priori e a posteriori. A informação a priori é um conhecimento prévio da associação entre as variáveis intermediárias (biomarcadores, nutrientes) e o desfecho (doença), e a posteriori é a combinação entre as variáveis intermediárias e o consumo alimentar (variáveis preditoras). A RRR tenta explicar o máximo possível da variação das variáveis intermediárias através das variáveis preditoras.

Objetivos: fornecer uma breve revisão teórica da técnica e descrever as rotinas computacionais em SAS.

Métodos: análise ilustrativa utilizando-se dados do estudo "Condições de saúde das mulheres: estudo de base populacional na região do Vale do Rio dos Sinos". Foram utilizadas como variáveis intermediárias o consumo dos nutrientes sódio, potássio e gordura saturada; as variáveis preditoras foram a frequência de consumo de 70 tipos de alimentos.

Conclusão: a RRR é uma poderosa ferramenta para detectar padrões alimentares que podem estar associados a alguma doença de interesse.

Palavras-chave: regressão por redução de posto (RRR), padrão alimentar, desfecho

Abstract

Background: the reduced rank regression (RRR) is a technique that has been used in nutritional epidemiology since 2004. Its goal is to find food patterns associated with a particular outcome. Thus, it is considered a technique which combines prior and posterior information. The prior information consists of a previous knowledge on the association between intermediate variables (biomarkers, nutrients) and outcome (disease). The posterior information consists of the combination between intermediate variables and food consumption (predictor variable).

Aims: to provide a brief theoretical review of the technique and to describe the computational routines in SAS software. Methods: An illustrative analysis using data from the study "Health conditions of women: a population-based study in the Vale do Rio dos Sinos" The intermediate variables were the consumption of the nutrients sodium, potassium, and saturated fat and the predictor variables were the frequencies of consumption of 70 foods.

Conclusion: The RRR is a powerful technique to detect food patterns that could be associated with a particular disease of interest.

Keywords: reduced rank regression (RRR), dietary pattern, outcome

Os métodos de regressão são as ferramentas estatísticas, que talvez sejam as mais amplamente usadas na análise de dados (1). As regressões lineares simples e múltipla, logística, Poisson e multivariada são alguns exemplos de modelos de regressão. A regressão é chamada de multivariada quando duas ou mais variáveis resposta são estudadas simultaneamente, ou seja, quando se deseja estudar duas ou mais variáveis dependentes ao mesmo tempo. Na literatura existe uma grande confusão com os termos "multivariada", "múltipla" e "multivariável" que seriam a tradução do inglês de multivariate, multiple e multivariable, respectivamente. Essa confusão existe tanto em português como em inglês. O mais adequado é utilizar o termo "multivariada" quando são utilizadas duas ou mais variáveis resposta e "múltipla" ou "multivariável" quando o modelo tem uma

única variável resposta e duas ou mais variáveis preditoras. A descrição usual do modelo de regressão multivariada é um conjunto de variáveis resposta relacionado a um conjunto de variáveis preditoras. Neste caso, temos então, que a matriz dos coeficientes de regressão é uma matriz de posto completo. Logo, quando o modelo possui muitas variáveis resposta e preditoras, temos que interpretar ao mesmo tempo um grande número de coeficientes de regressão, o que pode tornar-se uma tarefa bastante complicada. Então, em muitas situações, é necessário reduzir o número de parâmetros no modelo de regressão multivariada. O modelo de regressão por redução de posto (RRR) traz uma solução para essa questão. Mas a RRR não é útil apenas para reduzir o número de parâmetros, mesmo porque muitas vezes isso não ocorre, já que ela pode ser aplicada também para identificar uma relação entre as variáveis resposta, o que não existe na regressão multivariada.

Uma das áreas onde a redução de posto tem sido utilizada é a nutrição. Nesta área, uma das necessidades existentes é identificar padrões alimentares associados a algum tipo de doença. Geralmente, as técnicas estatísticas utilizadas para encontrar padrões alimentares são análise de componentes principais, análise fatorial e análise de agrupamento; sendo que estas focam apenas na identificação do padrão alimentar, sem ter como objetivo identificar padrões que estejam relacionados com alguma doença específica (2,3). A RRR, ao contrário dessas técnicas, busca identificar padrões alimentares relacionados a variáveis sabidamente associadas à doença.

Em geral, nessa área, as variáveis preditoras são o consumo de alimentos ou grupos de alimentos e as variáveis resposta, ou ainda variáveis intermediárias, são biomarcadores, consumo de nutrientes ou outras variáveis associadas com as variáveis preditoras e com o desenvolvimento da doença que estiver sendo estudada (desfecho). Assim, a técnica tenta explicar, através das variáveis preditoras, o máximo possível da variação das variáveis resposta (2).

O objetivo deste artigo é mostrar como aplicar e interpretar a RRR. Para atingi-lo organizamos o artigo da seguinte maneira: primeiramente, apresentamos um breve histórico do modelo e suas suposições; na sequência, mostramos a aplicação da RRR na nutrição apresentando as rotinas no SAS e a interpretação dos resultados. Não faremos nenhuma análise posterior a RRR para verificar a associação dos padrões encontrados com nenhum desfecho, pois os padrões aqui encontrados são apenas um exercício da técnica.

Regressão por redução de posto (RRR)

Em 1951, Anderson foi o primeiro a considerar o problema da regressão por redução de posto para os casos em que o conjunto de variáveis preditoras é fixo (1,4). Izenman em 1975 introduziu o termo Reduced Rank Regression (regressão por redução de posto - RRR) (5). De acordo com Reinsel (1), o modelo RRR e suas propriedades estatísticas foram estudados por diversos autores como Robinson (1973, 1974), Tso (1981), Davies e Tso (1982), Zhou (1994) e Geweke (1996). Anos mais tarde, em 2004, a técnica foi pela primeira vez aplicada na epidemiologia nutricional, por Hoffmann (2). A RRR também é conhecida por análise de redundância (6).

O modelo RRR é um caso especial da regressão multivariada, ou seja, para defini-lo basta tomar a equação definida por:

$$Y_k = CX_k + \varepsilon_k, k = 1,\ldots,T \quad (1)$$

e combiná-la com a restrição

$$posto(C) = r \leq \min(m,n). \quad (2)$$

Sendo que na equação (1) temos que:

$X_k = (x_{1k}, x_{2k}, \ldots, x_{nk})'$ é um vetor das n variáveis preditoras do k -ésimo indivíduo;

$Y_k = (y_{1k}, y_{2k}, \ldots, y_{mk})'$ é um vetor das m variáveis resposta do k -ésimo indivíduo;

$\varepsilon_k = (\varepsilon_{1k}, \varepsilon_{2k}, \ldots, \varepsilon_{mk})'$ é o vetor dos m erros aleatórios do k -ésimo indivíduo;

C é uma matriz $m \times n$ de coeficientes de regressão;

T é o número de indivíduos na amostra.

Assim, devido à restrição dada pela equação (2), C pode ser fatorada em $C = AB$, sendo que A tem dimensão $m \times r$ e B tem dimensão $r \times n$.

Então, podemos reescrever o modelo dado pela equação (1) como:

$$Y_k = A(BX_k) + \varepsilon_k, k = 1,\ldots,T \quad (3)$$

onde (BX_k) tem dimensão reduzida com r componentes.

Como no caso da nutrição, geralmente, temos mais variáveis preditoras do que variáveis resposta ($m < n$), se $r < \min(m,n)$ há uma redução no número de combinações lineares das variáveis preditoras para modelar a variação das variáveis resposta, pois serão necessárias r combinações lineares das variáveis preditoras ao invés de m combinações lineares (1).

A principal suposição da RRR é que tem distribuição normal multivariada (7). Na próxima seção mostraremos como testar essa suposição. Outra questão importante é determinar o número de fatores a serem extraídos. Existem várias maneiras de determinar esse número e ilustraremos um deles com o nosso exemplo.

Aplicação do modelo RRR

Origem dos Dados

Os dados que serão utilizados nessa seção foram cedidos pela coordenadora do projeto "Condições de saúde das mulheres: estudo de base populacional na região do Vale do Rio dos Sinos" (8), Maria Teresa Anselmo Olinto. Ao todo foram estudadas 1026 mulheres adultas de 20 a 60 anos residentes na zona urbana da cidade de São Leopoldo, RS, Brasil, que responderam um Questionário de Frequência Alimentar (QFA).

Descrição do Problema

O objetivo desta aplicação é identificar padrões alimentares em mulheres adultas residentes em São Leopoldo, RS, associados com hipertensão. Assim, para aplicarmos a RRR precisamos ter um conhecimento a priori da associação das variáveis resposta e o desfecho (2). Sabe-se que os nutrientes sódio e gordura saturada são fatores de risco para hipertensão e o nutriente potássio é fator protetor (9). E que o consumo desses nutrientes está associado com os alimentos ingeridos. Então, consideramos como desfecho a hipertensão arterial; como variáveis resposta os nutrientes sódio, potássio e gordura saturada e como variáveis preditoras a frequência de consumo de 70 tipos de alimentos.

RRR através do SAS

A seguir apresentaremos as rotinas do SAS para testar a normalidade multivariada e para fazer a RRR. Além do SAS, a RRR pode ser feita nos softwares R e S-PLUS utilizando uma sintaxe disponível em http://lib.stat.cmu.edu/S/rrr.s. Nas rotinas do SAS, os nutrientes sódio, gordura saturada e potássio são representados por SODIO, SATURADA, POTASSIO, respectivamente. Para representar o consumo dos alimentos foram utilizados os seguintes termos: AVEFARM2, PAOCENM2, PAOFANM2, PAOCASM2, LEITEINTM2, LEITEDESM2, LEITEFERM2, IOGURTEM2, QUEIJOM2, KASM2, ABACAM2, ABACAXIM2, BANANAM2, MAMAOM2, MACAM2, AMEIXAM2, CAQUVAM2, BERGAM2, LARANJAM2, LIMARM2, MANGAM2, MELAOM2, MORANGM2, ARROZINTM2, RROZBM2, MASM2, MASINM2, FEIJM2, FRANGOM2, PEIXEM2, GADOM2, PORCOM2, FIGADOM2, OVOSM2, PRESUNM2, LINGUIM2, ABOBORAM2, AGRIAOM2, ALHOM2, BATATAM2, AIPIMM2, BERINJM2, BROCOLISM2, COUVEM2, OUTVEGM2, SOJAM2, BANHAM2, CREMLEITM2, MAIOCASM2, MAIOINDM2, MANTM2, MARGM2, NATAM2, FRITM2, SOBRM2, SORVM2, CHOCOM2, BISDOCM2, BISSALM2, CUCAM2, AVELAM2, MCM2, IPOCAM2, MELM2, ACUCM2, ACUMASCM2, SUCNATM2, SUCINDM2, VINTINM2, CHOPM2.

• *Normalidade multivariada*

O teste da normalidade multivariada é feito através de uma macro que deve pode ser obtida em http://support.sas.com/kb/24/983.html. No mesmo endereço podem ser obtidos maiores detalhes sobre essa macro. Além da normalidade multivariada, ele testa a normalidade de cada variável separadamente. Primeiro exibiremos a sintaxe para realizar o teste e logo em seguida uma breve explicação do que cada comando faz. Os termos em negrito referem-se ao nosso exemplo:

```
1. %inc "F:\UFRGS\TCC\analises\multnorm.sas";
2. %multnorm(
      a. data=_proj_.banco_dados0,
      b. var=SODIO POTASSIO SATURADA,
      c. plot=mult)
```

1. Carrega a macro que testa a normalidade multivariada e que no nosso caso estava armazenada no endereço em negrito.

2. Executa a macro **multnorm** com os seguintes parâmetros:

a. conjunto de dados a ser utilizado;

b. variáveis que serão testadas quanto a normalidade multivariada;

c. solicita o gráfico do Quantil Qui-Quadrado versus o quadrado das distâncias Mahalanobis das observações.

• ***Regressão por redução de posto***

A rotina PROC PLS (6) do SAS 9.2 foi utilizada para fazermos a análise RRR. Novamente, primeiro mostraremos a sintaxe, a seguir a descrição de cada comando, utilizando negrito para destacar os termos específicos do nosso exemplo:

```
1. ods html;
2. ods graphics on;
3. proc pls
      a. data=_proj_.banco_dados0
      b. method=rrr
      c. varss
      d. details
      e. plots= ALL
      f. plots=(corrload(trace=off))
      g. cv=one
      h. cvtest;
4. model SODIO POTASSIO SATURADA = AVEFARM2 PAOCENM2 PAOFANM2 PAOCASM2
   LEITEINTM2 LEITEDESM2 LEITEFERM2 IOGURTEM2 QUEIJOM2 KASM2 ABACAM2
   ABACAXIM2 BANANAM2 MAMAOM2 MACAM2 AMEIXAM2 CAQUVAM2 BERGAM2
   LARANJAM2 LIMARM2 MANGAM2 MELAOM2 MORANGM2 ARROZINTM2 RROZBM2 MASM2
   MASINM2 FEIJM2 FRANGOM2 PEIXEM2 GADOM2 PORCOM2 FIGADOM2 OVOSM2
   PRESUNM2 LINGUIM2 ABOBORAM2 AGRIAOM2 ALHOM2 BATATAM2 AIPIMM2
   BERINJM2 BROCOLISM2 COUVEM2 OUTVEGM2 SOJAM2 BANHAM2 CREMLEITM2
   MAIOCASM2 MAIOINDM2 MANTM2 MARGM2 NATAM2 FRITM2 SOBRM2 SORVM2
   CHOCOM2 BISDOCM2 BISSALM2 CUCAM2 AVELAM2 MCM2 IPOCAM2 MELM2 ACUCM2
   ACUMASCM2 SUCNATM2 SUCINDM2 VINTINM2 CHOPM2;
5. output
      a. out=pattern
      b. xscore=scorex
      c. yscore=scorey
      d. stdy=ypad1-ypad3;
6. run;
7. ods graphics off;
8. ods html close;
```

Agora descreveremos cada comando.

1. Inicia a criação da saída em HTML.

2. Inicia a criação de gráficos que ilustram os resultados do procedimento pls.

3. Determina que o procedimento a ser utilizado é o pls, com os parâmetros a seguir:

a. conjunto de dados a ser utilizado;

b. método a ser utilizado é o rrr;

c. exibir percentual da variação explicada por cada fator para cada variável resposta e preditora;

d. exibir detalhes do modelo ajustado para cada fator;

e. construir todos os gráficos do procedimento pls;

f. retirar as linhas do gráfico de correlação das cargas (Correlation Loading Plot);

g. fazer validação cruzada do tipo one-at-a-time. Outras técnicas podem ser encontradas no site do SAS (6);

h. utilizar teste de Van der Voet para determinar quantos fatores devem ser extraídos. Para maiores detalhes ver: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_pls_sect014.htm

4. Indica o modelo a ser ajustado sendo que as variáveis descritas antes do sinal de igualdade são as variáveis resposta e as localizadas após o sinal são as variáveis preditoras.

5. Define a criação de um arquivo que conterá as variáveis definidas a seguir:

a. nome do banco que conterá os dados originais e as variáveis criadas na sequência;

b. nome das variáveis que receberão os valores dos escores das variáveis preditoras cada fator extraído, portanto o número de variáveis criadas é igual ao número de fatores extraídos. No nosso exemplo os nomes serão scorex1, scorex2, etc;

c. idem item (5b) mas para os escores das variáveis resposta. Esses escores serão utilizados para comparar os sujeitos com e sem a doença de interesse;

d. nome das variáveis que receberão os valores das variáveis resposta padronizadas. Aqui, ao contrário dos itens (5a) e (5b) é necessário especificar o nome de cada variável que será criada. No nosso exemplo, foram criadas as variáveis ypad1, ypad2 e ypad3.

6. Executa os comandos acima.

7. Finaliza a criação dos gráficos.

8. Finaliza a criação da saída em HTML.

Resultados

Nesta subseção descreveremos os resultados produzidos pelas sintaxes explicadas anteriormente. Eles são exibidos exatamente como o SAS os produz, a não ser por uma numeração que foi acrescentada ao lado direito das tabelas e/ou nos títulos das colunas. Ela foi acrescentada para facilitar a interpretação dos resultados que será descrita após as saídas do SAS.

• *Teste de normalidade multivariada*

MULTNORM macro: Univariate and Multivariate Normality Tests

The MODEL Procedure

	Normality Test			
	Equation	Test Statistic	Value	Prob
1.	SODIO	Shapiro-Wilk W	0.96	<.0001
2.	POTASSIO	Shapiro-Wilk W	0.98	0.0589
3.	SATURADA	Shapiro-Wilk W	0.98	<.0001
4.	System	Mardia Skewness	85.88	<.0001
5.		Mardia Kurtosis	0.61	0.5429
6.		Henze-Zirkler T	9.26	<.0001

A primeira parte consiste do teste de normalidade de Shapiro-Wilk para cada uma das variáveis:

1. SODIO - Rejeita-se a hipótese de normalidade para SODIO (Shapiro-Wilk W= 0,96; p<0,0001).

2. POTASSIO - Rejeita-se a hipótese de normalidade para POTASSIO (Shapiro-Wilk W= 0,98; p=0,0589).

3. SATURADA - Rejeita-se a hipótese de normalidade para SATURADA (Shapiro-Wilk W= 0,98; p<0,0001).

A segunda parte refere-se aos testes multivariados.

4. Teste de Mardia para o coeficiente de assimetria da normal multivariada. Rejeita-se a hipótese de assimetria normal multivariada (Mardia Skewness= 85,88; p<0,0001).

5. Teste de Mardia para o coeficiente de curtose da normal multivariada. Aceita-se a hipótese de curtose normal multivariada (Mardia Kurtosis= 0,61; p= 0,5429).

6. Teste de Henze-Zirkler T para testar a normalidade multivariada. Rejeita-se a hipótese de normalidade multivariada (Henze-Zirkler T= 9,26; p<0,0001).

A Figura 1 apresenta o gráfico do quantil-quantil qui-quadrado do quadrado das distâncias de Mahalanobis das observações em relação ao vetor das médias. Para um número grande de variáveis e tamanho de amostra grande é esperado que os quadrados das distâncias de Mahalanobis das observações em relação ao vetor das médias tenha distribuição Qui-quadrado (10), portanto quanto mais afastados os pontos estiverem da reta que representa a identidade, maior a chance dos dados não terem distribuição multivariada.

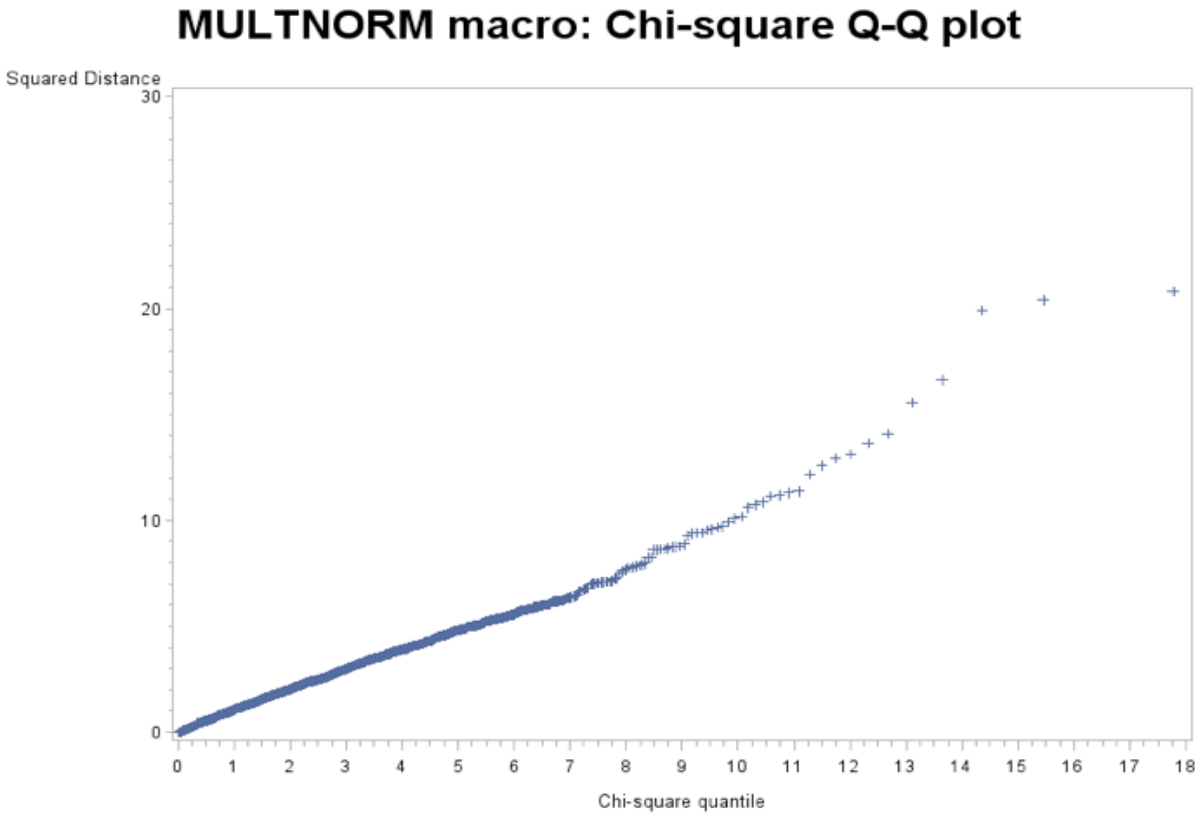


Figura 1: Quantil-quantil qui-quadrado versus o quadrado da distância de Mahalanobis.

Apesar do teste rejeitar a normalidade multivariada das variáveis resposta, percebemos pela Figura 1 que não há um grande desvio da normalidade. A mesma figura confirma o que o teste de Mardia para os coeficiente de assimetria e curtose da normal multivariada já haviam apontado: o problema é de assimetria e não de curtose. Dessa forma, procederemos a análise sem fazer nenhuma transformação nas variáveis.

• *Análise por redução de posto (RRR)*

As saídas apresentadas a seguir referem-se apenas à descrição do banco de dados utilizado e da análise aplicada. Eles serão explicados pelos itens de 1 a 12.

RRR

	The PLS Procedure	
1.	Data Set	_PROJ_.BANCO_DADOS0
2.	Factor Extraction Method	Reduced Rank Regression
3.	Number of Response Variables	3
4.	Number of Predictor Parameters	70
5.	Missing Value Handling	Exclude
6.	Maximum Number of Factors	3
7.	Validation Method	Leave-one-out Cross Validation
8.	Validation Testing Criterion	Prob T**2 > 0.1
9.	Number of Random Permutations	1000
10.	Random Permutation Seed	903883001

As linhas dessa tabela representam:
1. Nome do banco de dados utilizado.
2. Método utilizado.
3. Número de variáveis resposta utilizadas.
4. Número de variáveis preditoras utilizadas.
5. Método utilizado para lidar com os dados faltantes.
6. Número máximo de fatores que podem ser extraídos.
7. Método de validação cruzada utilizado.
8. Critério utilizado para determinar quantos fatores serão extraídos.
9. Número de permutações aleatórias utilizado.
10. Semente para permutação aleatória utilizada.

11.	Number of Observations Read	1026
12.	Number of Observations Used	1023

11. Número de observações do banco de dados.
12. Número de observações utilizado na análise, ou seja, excluídos os dados faltantes.

As próximas saídas referem-se aos resultados da validação cruzada e são explicadas pelos itens de 13 a 16 e pela figura 2.

	Cross Validation for the Number of Extracted Factors			
	Number of Extracted Factors	Root Mean PRESS	T**2	Prob > T**2
13.	0	1.000978	475.7195	<0.0001
	1	0.557117	512.1546	<0.0001
	2	0.337668	329.8757	<0.0001
	3	1.03E-14	0	1.0000

14.	Minimum root mean PRESS	1.03E-14
15.	Minimizing number of factors	3
16.	Smallest number of factors with p > 0.1	3

13. Em cada linha é exibido o valor da raiz quadrada da média da soma de quadrados do resíduo (root mean PRESS), a estatística T2 de Hotelling e o seu respectivo valor p para cada um dos fatores. Como todo teste de ajuste, o que esperamos aqui é que a hipótese nula não seja rejeitada, ou seja, queremos valores grandes de p.

14. O root mean PRESS mínimo, que também pode ser utilizado como critério de escolha do número de fatores.

15. Número de fatores em que o root mean PRESS atinge seu valor mínimo.

16. Menor número de fatores com p>0,1. Neste caso, é usada a estatística T2 de Hotelling como critério de escolha do número de fatores.

Como estamos usando a opção plots=ALL, são gerados todos os gráficos associados aos resultados. Desse modo, o pesquisador pode optar por exibir os resultados na forma de texto, de tabelas ou de gráficos. Assim sendo, o gráfico na parte superior da Figura 2 é apenas a ilustração das informações apresentadas anteriormente (item 13). Já os dados utilizados na parte inferior da figura 2 constam na saída Percent Variation Accounted for by Reduced Rank Regression Factors apresentada a seguir. Nessa figura, note que a escala do eixo Y está em percentual. A linha verde tracejada (Model Term R Square) representa a proporção da variação das variáveis preditoras explicada por um dois ou três fatores. A linha marrom tracejada (Dependent Variable R Square) representa a proporção da variação das variáveis resposta explicada por um dois ou três fatores.

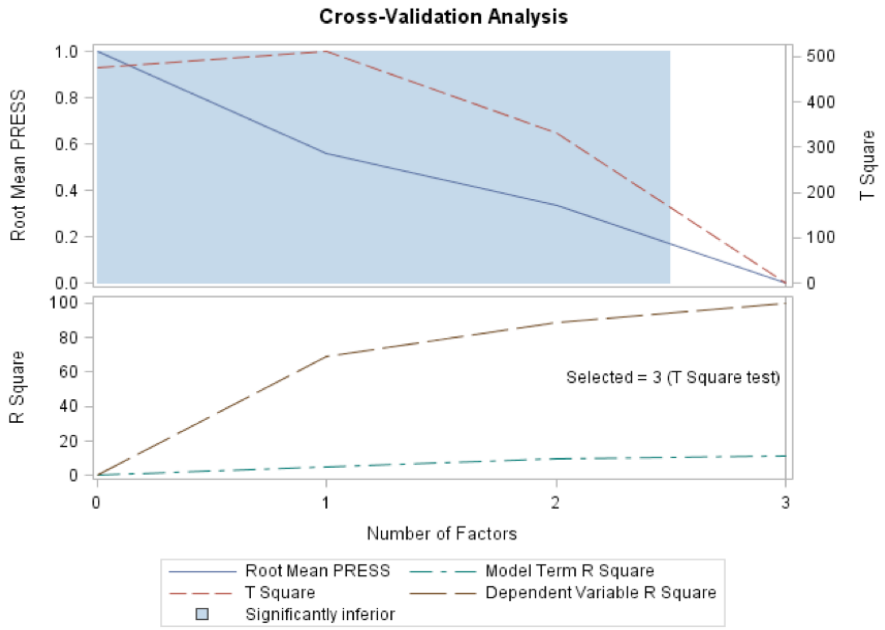


Figura 2: Gráfico de validação cruzada.

Na saída a seguir é apresentado o percentual de variação explicada pelos fatores obtidos pela RRR. Como a saída originalmente gera uma tabela com mais de 70 colunas, vamos exibir aqui apenas os resultados do consumo de aveia (AVEFARM2), omitindo todos os demais itens alimentares. A saída completa pode ser obitda em http://www.mat.ufrgs.br/~camey/RRR_nutrientes/RRR_nutrientes.htm.

Percent Variation Accounted for by Reduced Rank Regression Factors

	Number of Extracted Factors	Model Effects				Dependent Variables				
		a	b	c	D	e			f	g
		AVEFARM2	...	Current	Total	SODIO	POTASSIO	SATURADA	Current	Total
	1	1.8139	...	5.2885	5.2885	76.5201	57.8701	72.8369	69.0757	69.0757
17.	2	8.5469	...	4.0816	9.3701	81.4232	99.6234	85.0176	19.6123	88.6880
	3	8.5564	...	1.6218	10.9919	100.0000	100.0000	100.0000	11.3120	100.0000

13. Cada linha representa o percentual de variação explicada pelo conjunto dos fatores (com excessão das colunas (c) e (f)) para:

a. **Aveia**. Podemos interpretar os números dessa coluna da seguinte maneira: 1,8% da variação do consumo de aveia é explicada pelo fator 1; os fatores 1 e 2 conjuntamente explicam 8,5% da variação do consuno de aveia; e finalmente, os três fatores explicam 8,6% da variação do consumo de aveia.

b. Apenas para lembrar que o consumo dos demais alimentos foi suprimido.

c. **Conjunto de todos os alimen**tos. Aqui a intepretação é diferente, pois não expressa o percentual de variação explicada pelo conjunto de fatores, mas sim de cada fator. Dessa forma, temos que o primeiro fator explica 5,3% da variação de todas as variáveis preditoras; o segundo fator explica 4,1% e o terceiro fator explica 1,6% da variação de todas variáveis preditoras.

d. **Conjunto de todos os alimentos**. Pode-se notar que essa coluna é apenas a soma acumulada da coluna (c). Resumidamente, podemos concluir que os três fatores explicam conjuntamente 11,0% da variação de todas as variáveis preditoras.

e. **Cada um dos nutrientes**. Nas três colunas a seguir temos os percentuais de variação explicada do consumo dos nutrientes. A interpretação é a mesma da coluna (a).

f. **Conjunto de todos os nutrientes**. Idem coluna (c).

g. **Conjunto de todos os nutrientes**. Idem coluna (d).

Aqui cabe uma observação de que os consumos de sódio, potássio e gordura saturada foram calculados a partir da frequência de consumo dos 70 itens alimentares. Por essa razão o modelo atinge 100% de explicação. Podemos também concluir que o fator 3 pouco contribui para a

explicação da variação do consumo dos alimentos e dos nutrientes, apenas 1,6% e 11,3%, respectivamente. Neste caso, poderíamos optar por extrair apenas dois fatores. Devemos também lembrar que os fatores são ordenados por ordem de explicação, isto é, o fator 1 é aquele que mais explica a variação e o fator 3 é o que menos explica.

As próximas duas saída trazem as cargas fatoriais (loadings) e os pesos (weights), respectivamente, de cada alimento em cada fator. A saída com as cargas fatoriais é muito útil para verificar a importância de cada alimento dentro de cada fator. Quanto maior o valor absoluto da carga, maior é a importância da variável para aquele fator. O sinal da carga indica a direção da associação da frequência de consumo do alimento com o fator. Por exemplo, a carga da variável aveia no fator 1 é 0,07; no fator 2 é 0,15 e no fator 3 é 0,01. Em relação aos dois elementos exibidos aqui, podemos ver que a aveia tem mais influência no fator 2 que o chop. Além disso, eles têm direções opostas, ou seja, quanto maior a frequência de consumo de aveia, maior o escore do fator 2; ao contrário da frequência de consumo do chop, pois quanto maior for o consumo, menor o escore desse fator. A saída com os pesos só tem utilidade se quisermos exibir como calcular o escore de cada fator a partir da frequência de consumo dos alimentos.

Model Effect Loadings

Number of Extracted Factors	AVEFARM2	...	CHOPM2
1	0.069998	...	0.043374
2	0.153512	...	-0.033164
3	0.009148	...	-0.003358

Model Effect Weights

Number of Extracted Factors	AVEFARM2	...	CHOPM2
1	0.049953	...	0.004494
2	0.070321	...	0.005322
3	0.013203	...	-0.000366

A próxima saída nos fornece os pesos das variáveis resposta dentro de cada fator. Temos então que o peso da variável SODIO é igual a 0,61 no fator 1; no fator 2 igual a -0,29 e no fator 3 igual a -0,74; o peso da variável POTASSIO é igual a 0,53 no fator 1; 0,84 no fator 2 e 0,11 no fator 3; e finalmente, o peso da variável SATURADA é igual a 0,59 no fator 1; -0,45 no fator 2 e 0,66 no fator 3. Esses valores serão utilizados para calcular os escores dos fatores.

Number of Extracted Factors	Dependent Variable Weights		
	SODIO	POTASSIO	SATURADA
1	0.607665	0.528450	0.592860
2	-0.288674	0.842403	-0.454999
3	-0.739872	0.105344	0.664449

Como vimos anteriormente podemos calcular os escores de cada fator tanto pelos alimentos quanto pelos nutrientes. Ou seja, a RRR nos fornece uma ligação entre alimentos -> fatores e nutrientes -> fatores. Mas, ela também fornece uma ligação direta entre alimentos -> nutrientes. Essa ligação é obtida a partir dos coeficientes exibidos nas próximas três saídas. A partir deles podemos estimar o consumo de cada nutriente a partir do consumo dos alimentos. O que difere as três saídas é que a primeira delas traz os coeficientes supondo que apenas 1 fator foi extraído, a segunda supõe que os dois primeiros fatores foram extraídos e a terceira supõe que todos os fatores foram extraídos.

Coded Regression Coefficients for 1 Extracted Factor

	SODIO	POTASSIO	SATURADA
AVEFARM2	0.0303548244	0.0263977649	0.0296152657
...	...	...	...
CHOPM2	0.0027309074	0.0023749059	0.0026643721

Coded Regression Coefficients for 2 Extracted Factors

	SODIO	POTASSIO	SATURADA
AVEFARM2	0.0100550609	0.0856361458	-0.0023805675
...	...	...	...
CHOPM2	0.0011945814	0.0068581830	0.0002428645

Coded Regression Coefficients for 3 Extracted Factors

	SODIO	POTASSIO	SATURADA
AVEFARM2	0.0002862533	0.0870270377	0.0063924081
...	...	...	
CHOPM2	0.0014650138	0.0068196786	0.0000000000

A Figura 3 traz a correlação das cargas fatoriais dos dois primeiro fatores. A partir dela podemos extrair as seguintes informações:

• Encontrar padrões e agrupamentos dos indivíduos (cada indivíduo tem um número e ele é colocado no gráfico em cor verde). Se mais de um agrupamento for visualizado nesse gráfico, então pode ser necessário identificar os grupos e realizar a análise para cada um dos grupos. No nosso exemplo podemos ver claramente um único agrupamento dos indivíduos.

• Ver quantidade de explicação da variação da variável resposta. Assim, quanto maior for a explicação de cada

variável resposta (cor vermelha), mais próxima do círculo 100% ela estará.

• As cargas mostram o quanto da variação de cada variável é explicada pelos dois primeiros fatores conjuntamente pela distância da variável até a origem (onde a variável se encontra no círculo) e individualmente pelas projeções dessa variável sobre o eixo horizontal e vertical. Portanto, podemos ver que os fatores 1 e 2 explicam aproximadamente 75% da variação (o valor exato pode ser obtido nas tabelas das saídas do SAS) do sódio e da gordura saturada, sendo que o fator 1 explica muito mais da variação desses nutrientes, uma vez que projetando esses pontos nos eixos X e Y temos em torno de 0,75 (75%) e -0,25 (25%), respectivamente. Esse mesmo tipo de interpretação pode ser feito para os alimentos.

• Analisar a relação das variáveis resposta e preditoras. Por exemplo, a variável PRESUNM2 (presunto) está altamente relacionada às variáveis resposta SATURADA e SODIO.

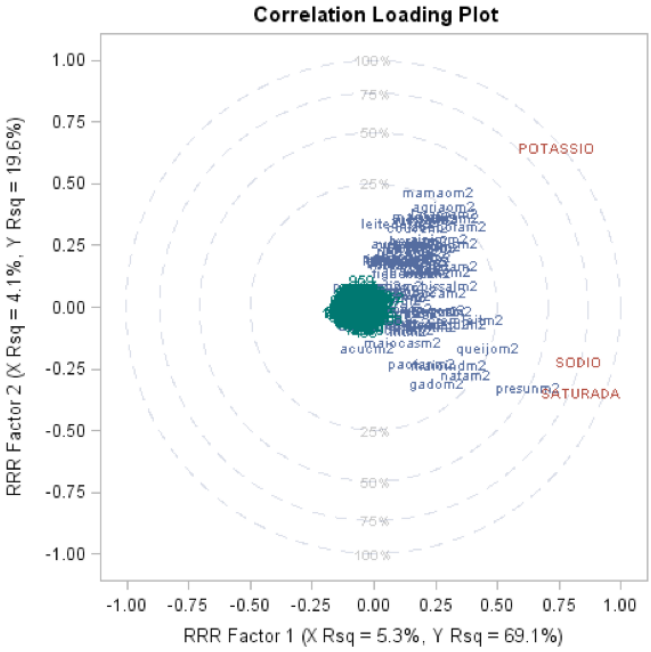


Figura 3: Gráfico de correlação das cargas.

A Figura 4 mostra a matriz de gráficos dos escores de cada fator obtido através das variáveis preditoras. Na diagonal principal, tem-se os histogramas dos escores de cada fator. Nas demais posições, tem-se gráficos de dispersão dos escores fatoriais dois a dois. Por exemplo, na linha 1 e coluna 1, temos os escores das variáveis preditoras do fator 1 versus os do fator 2, não encontramos nenhum padrão neles (nuvem), mostrando a independência entre os fatores. O mesmo gráfico é fornecido para os escores fatoriais obtidos a partir das variáveis resposta.

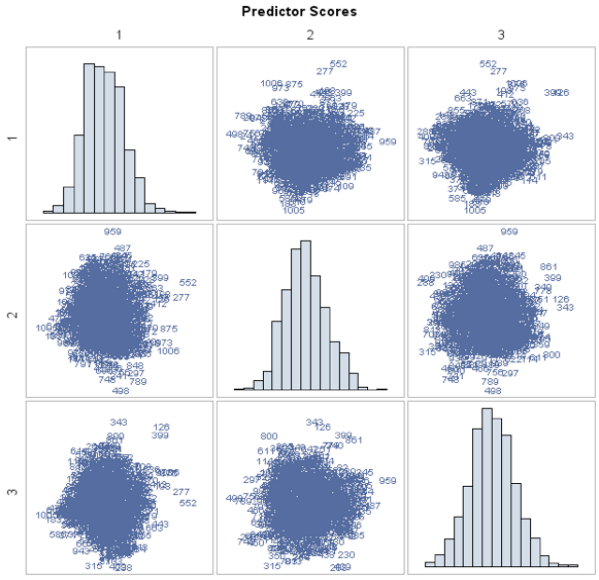


Figura 4: Matriz de gráficos dos escores das variáveis preditoras de cada fator.

A Figura 5 mostra os escores fatoriais das variáveis resposta versus os escores fatoriais das variáveis preditoras de cada fator. Nesse exemplo, existe alta correlação entre os escores de X e Y para os fatores 1, 2 e 3. Isso se deve ao fato, já mencionado, das variáveis resposta terem sido calculadas a partir das variáveis preditoras.

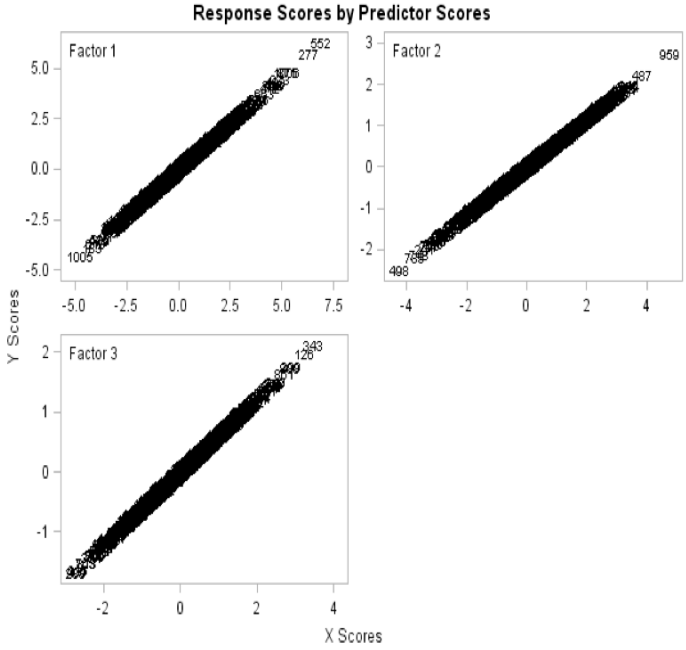
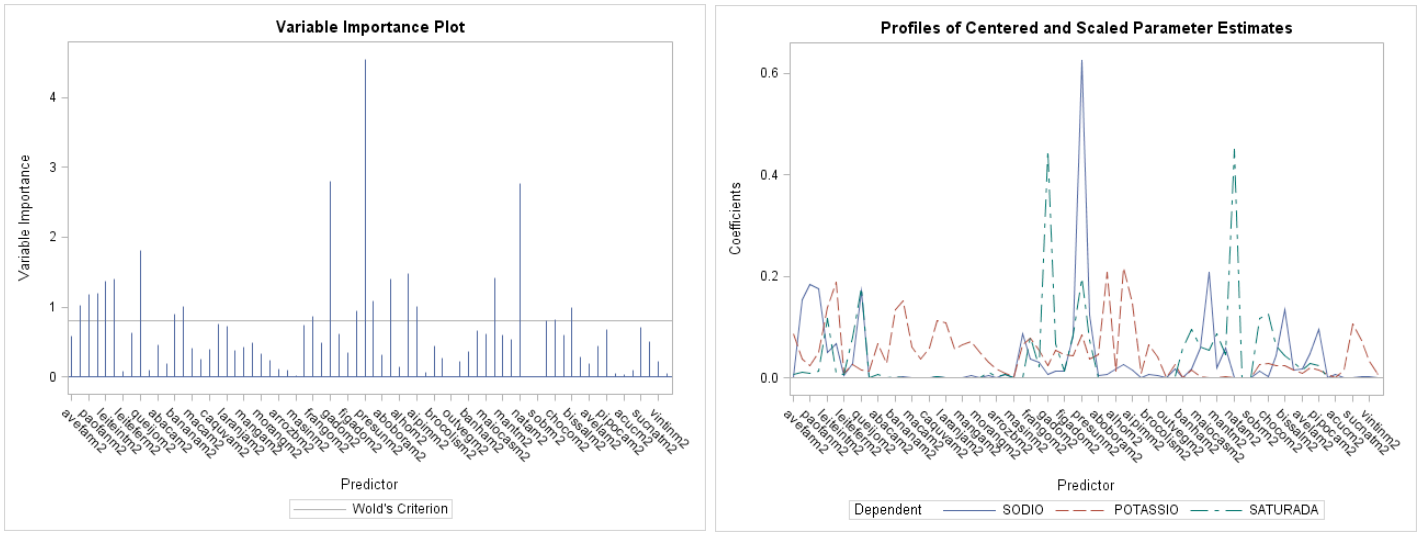


Figura 5: Gráficos dos escores das respostas versus os escores das preditoras.

A Figura 6 (a) representa a contribuição de cada variável preditora no modelo RRR de uma forma geral. Essa importância é medida pelo que o SAS chama de "importância da variável" (VIP). Já a figura 6 (b) mostra a importância de cada preditora na estimação de cada variável resposta. Assim, se uma variável preditora tem um baixo valor de VIP (Critério de Wold: VIP < 0,8 – linha cinza horizontal no gráfico) e um coeficiente pequeno (em valor absoluto) então ela é uma forte candidata a ser retirada do modelo.

Como são muitos itens alimentares, os gráficos exibem os nomes alternadamente. Por exemplo, o último nome exibido é vintinm2, mas o último alimento é o chopm2, que não é exibido. Portanto, com relativa facilidade podemos determinar quais são os itens que poderiam ser retirados da análise. No nosso exemplo, podemos notar que o Chop (chopm2) é um candidato a ser retirado, pois ele tem baixo valor de VIP e valores baixos de coeficientes para as três variáveis resposta.



a b

Figura 6: (a) Gráfico de importância das variáveis (VIP).
(b) Perfil das estimativas dos parâmetros centrados e escalonados.

A figura 7 é uma maneira de ver quais itens mais influenciam em cada um dos fatores, bem como a direção da associação. Os valores usados para construir esse gráfico são provenientes da saída Model Effect Loadings. Se considerarmos apenas os itens com carga maior do que 0,2, em módulo, temos que os itens mais relevantes para o fator 1 são: queijo (queijom2), presunto (presunm2) e nata (natam2).

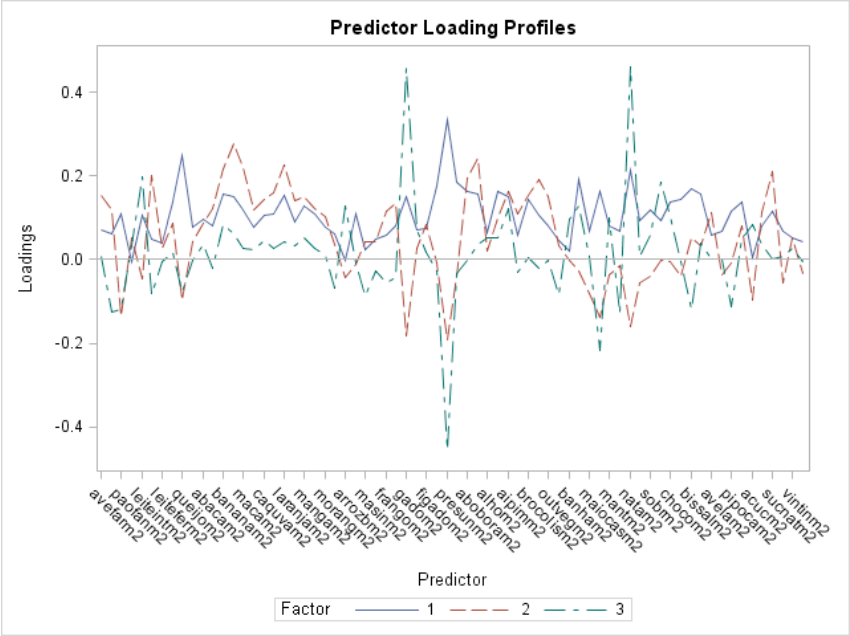


Figura 7: Perfil dos pesos das variáveis preditoras em cada um dos fatores.

Os próximos gráficos são úteis para dectectar outliers. O gráfico da figura 8 mostra a distância de cada observação até o modelo para as variáveis resposta. Assim, o modelo não é adequado para as observações que têm distâncias muito grandes quando comparadas as demais. O problema desse gráfico é que ele não permite identificar qual é observação que está muito distante do modelo. Ele é mais útil para ter uma visão geral do comportamento das observações. Existe uma versão semelhante desse gráfico para o modelo para as preditoras.

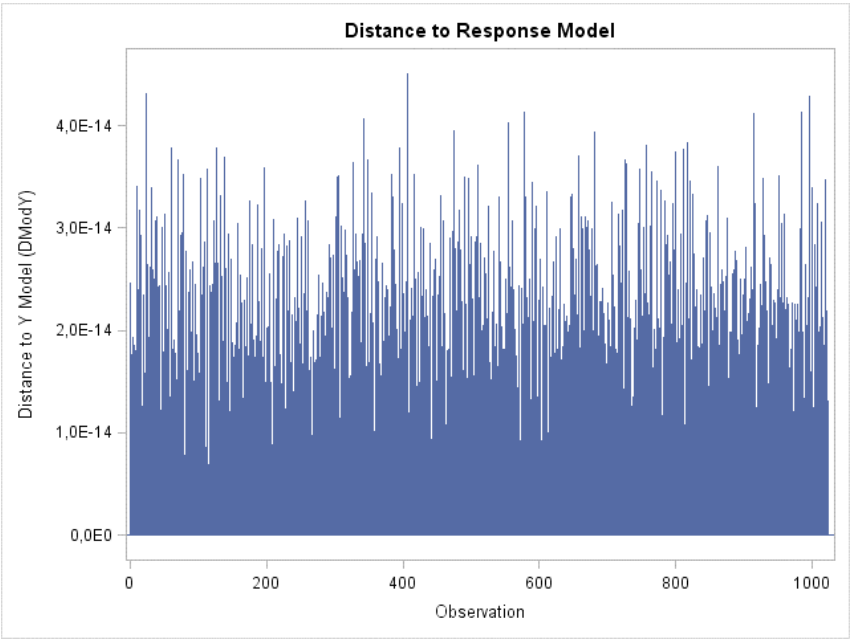


Figura 8: Distância de cada observação até o modelo para as variáveis resposta.

A figura 9 fornece um gráfico de dispersão entre a distância das observações até o modelo para as variáveis resposta e para as preditoras. A sua principal vantagem é permitir visualizar quais são as observações que estão muito distantes dos modelos. Por exemplo, a observação 126 tem uma distância muito grande em relação ao modelo para as preditoras e para as variáveis resposta, ou seja, poderíamos então retornar aos dados e tentar entender se há uma explicação para essa observação apresentar esse comportamento. Se for possível identificar que se trata de um valor extremo, ou seja, que não é o esperado da população em estudo, poderíamos retirar essa observação e refazer a análise. Outro exemplo é a observação 409, mas, neste caso, ela se distancia apenas do modelo das preditoras. Mesmo assim poderíamos fazer a mesma análise e tomar a mesma ação descrita para a observação 126. Devemos aqui relembrar que a RRR prioriza a explicação da variância das variáveis resposta (Y) e, por essa razão, o modelo apresenta um ajuste melhor para estas variáveis do que para as variáveis preditoras (X). Esse fato pode ser notado através das diferenças nas escalas dos eixos X e Y.

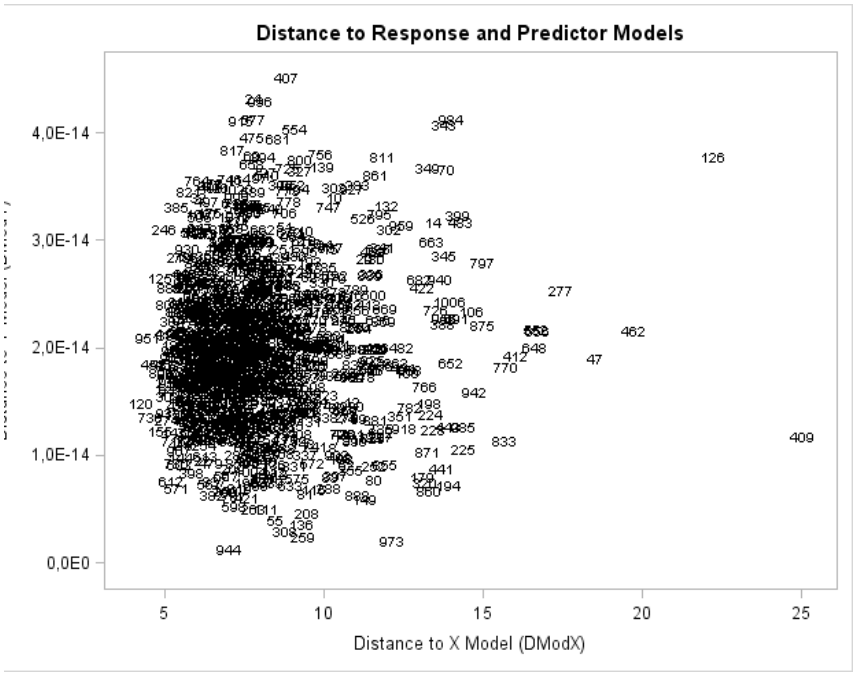


Figura 9: Distância das observações até o modelo para as resposta versus a distância das observações até o modelo para as preditoras.

As duas próximas figuras são úteis para analisar a qualidade do ajuste do modelo. A figura 10 traz os gráficos de diagnóstico de ajuste do modelo para a variável POTASSIO. O programa também gera o mesmo tipo de gráfico para as variáveis SODIO e SATURADA, mas eles serão omitidos. A figura 10 (i) é o gráfico dos valores preditos versus os resíduos. Ele apresenta um comportamento errático, ou seja, os resíduos não têm nenhum padrão definido. Isso indica uma boa qualidade do ajuste. A figura 10 (ii) traz o histograma dos resíduos com curva da distribuição. Através dele, podemos verificar se os resíduos têm aproximadamente distribuição normal, o que é uma suposição do modelo, e parece acontecer nesse caso. A figura 10 (iii) exibe o gráfico de probabilidade normal, que também é útil para verificar a suposição de normalidade dos resíduos. Quanto mais próximos os pontos estiverem da linha diagonal traçada melhor. A figura 10 (iv) é chamada de Residual-Fit ou RF Plot. Nesse gráfico podemos avaliar se o modelo é inadequado ou não. Se a dispersão dos resíduos for maior que a dispersão do ajuste centrado, o modelo é considerado inadequado. A figura 10 (v) é o gráfico da variável resposta POTASSIO versus valores preditos. Ele confirma os resultados dos gráficos anteriores, que o modelo ajustado faz boas predições (valores em torno da linha diagonal). A figura 10 (iv) é o gráfico dos resíduos absolutos versus valores preditos. Tem uma interpretação semelhante ao do gráfico (i), porém os resíduos negativos são projetados na parte positiva.

O programa gera gráficos dos resíduos de cada variável resposta versus valores observados de cada uma das variáveis preditoras, vamos exibir aqui apenas alguns deles para o potássio (figura 11). Nos quatro gráficos exibidos na figura 11 podemos ver que não há nenhum padrão, indicando que o modelo parece estar bem ajustado.

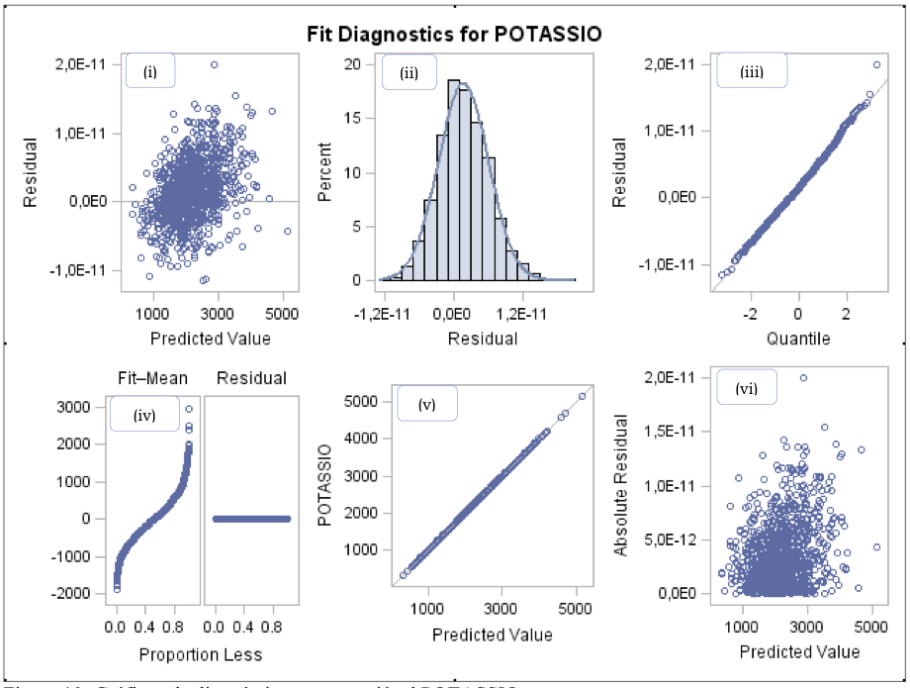


Figura 10: Gráficos de diagnóstico para a variável POTASSIO.

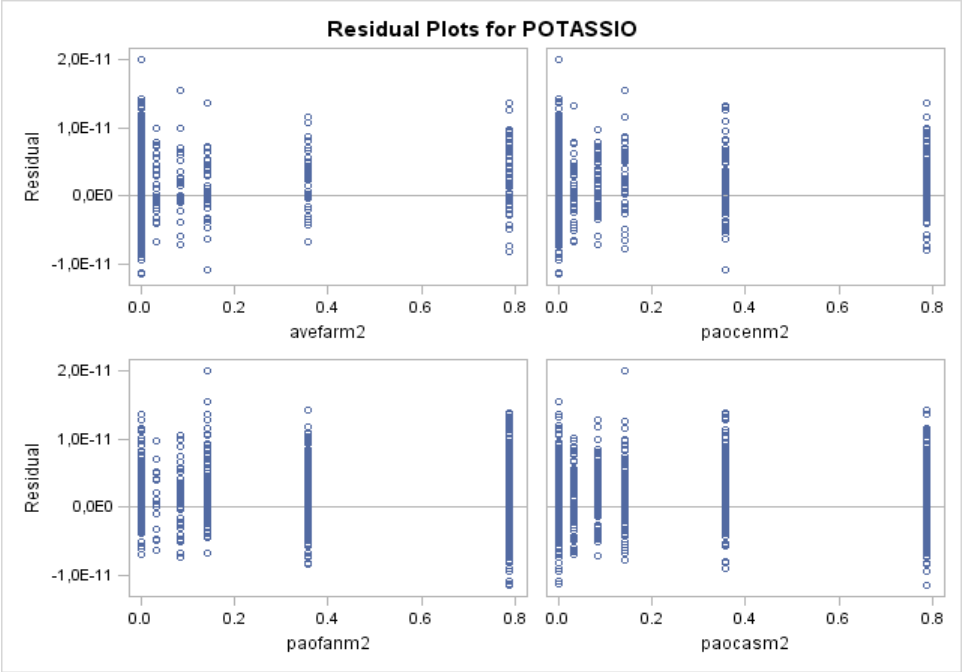


Figura 11: Gráfico dos resíduos versus variáveis preditoras para POTASSIO.

Interpretação dos fatores

É importante sabermos como os escores das variáveis resposta são calculados, pois isso facilita a interpretação dos fatores. A seguir exemplificaremos o cálculo para o sujeito 1. Primeiro temos que padronizar o consumo de cada nutriente, ou seja, calcular o escore Z de cada um. Para o sujeito 1 temos que os escores Z de cada nutriente são:

$$Z_{SODIO} = \frac{1851,71 - 1388,22}{528,44} = 0,877$$

$$Z_{POTASSIO} = \frac{3088,08 - 2192,61}{700,05} = 1,279$$

$$Z_{SATURADA} = \frac{35,37 - 28,63}{10,73} = 0,628$$

Podemos interpretar que esse indivíduo, portanto, ingere acima da média sódio, potássio e gordura saturada, sendo que consume mais do que um desvio-padrão acima da média do consumo de potássio. Os valores das médias (DP) do sódio, potássio e gordura saturada são, respectivamente, 1388,22 (528,44); 2192,61 (700,05) e 28,63 (10,73).

Tendo os escores Z de cada um dos nutrientes podemos calcular os escores fatoriais do sujeito 1 da seguinte forma:

Fator	Peso da variável SODIO no fator	Valor de SODIO padronizada	Peso da variável POTASSIO no fator	Valor de POTASSIO padronizada	Peso da variável SATURADA no fator	Valor de SATURADA padronizada	Escore do fator
1	0,608 ×	0,877 +	0,528 ×	1,279 +	0,593 ×	0,628 =	1,581
2	-0,289 ×	0,877 +	0,842 ×	1,279 +	-0,455 ×	0,628 =	0,539
3	-0,740 ×	0,877 +	0,105 ×	1,279 +	0,664 ×	0,628 =	-0,097

Os pesos de cada variável em cada fator são fixos para todos os sujeitos (ver saída Dependent Variable Weights). Como os pesos das três variáveis são positivos temos que o escore do fator 1 aumentará a medida que o consumo de sódio, potássio e gordura saturada crescer em relação a média. Já no fator 2 apenas o potássio tem peso positivo; portanto, o escore do fator 2 será alto para indivíduos que consomem potássio acima da média e sódio e gordura saturada abaixo da média. E o escore do fator 3 será alto para indivíduos que consomem gordura saturada acima da média e sódio abaixo da média. Fazendo uma interpretação grosseira, poderíamos dizer que o fator 1 estaria ligado a uma dieta rica nos 3 nutrientes, o fator 2 a uma dieta mais saudável, ou seja, rica em potássio e pobre em sódio e gordura saturada e, finalmente, o fator 3 estaria associada a uma dieta rica em gordura saturada, mas pobre em sódio. Quando associamos essa interpretação com os resultados exibidos na figura 7, temos que o fator 1 também está associado ao consumo de queijo (queijom2), presunto (presunm2) e nata (natam2).

Os escores fatoriais aqui obtidos podem ser utilizados como um escore contínuo ou categorizado (divididos em tercis, quartis, quintis) e usados:

• para comparar suas médias entre o grupo de pacientes com ou sem a doença;

• para entrar como covariável em regressões logística, Poisson ou linear (dependendo de como é tratado o desfecho e do tipo de delineamento utilizado).

Conclusão

Neste trabalho, a Regressão por Redução de Postos (RRR) foi estudada e exemplificada no contexto de epidemiologia nutricional. Foi visto que o método RRR combina informação a priori e a posteriori, ou seja, é a priori, pois utiliza a informação existente da associação entre o desfecho e as variáveis intermediárias e é a posteriori, pois utiliza a informação dos dados do estudo de consumo alimentar.

No entanto, se não houver a informação a priori da associação das variáveis intermediárias com o desfecho, devemos preferir a técnica exploratória de análise de componentes principais (PCA) à RRR, pois sem esse conhecimento a priori não podemos justificar o uso dessas variáveis resposta (2).

Como visto nesse trabalho, uma das suposições do modelo é que as variáveis resposta tenham distribuição normal multivariada, por isso é importante que seja feito o teste de normalidade multivariada para as variáveis resposta. E também precisamos ser cuidadosos em relação aos dados faltantes das variáveis resposta e das variáveis preditoras, uma vez que a RRR trabalha com matrizes de completas, ou seja, se houver algum dado faltante nas variáveis resposta e/ou nas variáveis preditoras de um indivíduo, esse sujeito não entrará na análise; logo, perderemos informações. Também é fundamental que seja feita uma boa análise de resíduos para investigarmos a adequabilidade do modelo ajustado.

É importante ressaltar que na RRR, o número máximo de fatores que a técnica encontra é igual ao número de variáveis resposta existentes (2) , o que ocorreu no exemplo considerado. No entanto, podemos ter a redução do número de fatores, para isso precisamos fazer um teste para saber qual é o número de fatores ideal para ser utilizado na análise ou avaliar qualitativamente a importância de cada fator.

Uma desvantagem da RRR apontada por Hoffmann (2004) é que os coeficientes dos escores dos fatores são estimados através dos dados disponíveis e eles não podem ser reproduzidos para dados de outra população que estiver sendo estudada.

Alertamos também que nesse trabalho foi utilizado o aplicativo SAS, mas a RRR pode ser feita nos aplicativos R e S-PLUS através da macro disponível em http://lib.stat.cmu.edu/S/rrr.s.

Referências

1. Reinsel G, Velu RP. Multivariate reduced-rank regression : theory and applications. New York: Springer; 1998. 258 p.
2. Hoffmann K, Schulze MB, Schienkiewitz A, Nöthlings U, Boeing H. Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. American Journal of Epidemiology. 2004;159:935–44.
3. Schulze MB, Hoffmann K. Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. Br. J. Nutr. 2006;95(5):860–9.
4. Aldrin M. Reduced-Rank regression. Encyclopedia of environmetrics. 2001;3:1724–8.
5. Izenman A. Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis. 1975 Jun;5:248–64.
6. The PLS Procedure: Overview: SAS/STAT(R) 9.2 User's Guide, Second Edition [Internet]. [cited 2011 Nov 10]; Available from: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_pls_sect001.htm
7. Timm N. Applied multivariate analysis. New York: Springer,; 2002.
8. Alves ALS, Olinto MTA, Costa JSD da, Bairros FS de, Balbinotti MAA. Padrões alimentares de mulheres adultas residentes em área urbana no sul do Brasil. Revista de Saúde Pública [Internet]. 2006 Oct [cited 2011 Oct 21]; 40. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102006000600017&lng=pt&nrm=iso&tlng=p
9. 11752.pdf [Internet]. [cited 2011 Oct 21]; Available from: http://www.scielo.br/pdf/abem/v43n4/11752.pdf
10. 24983 - Macro to test multivariate normality [Internet]. [cited 2011 Nov 10]; Available from: http://support.sas.com/kb/24/983.html

Recebido: 21/11/2011
Aceito: 23/12/2011